A Metric for Evaluating and Comparing Hierarchical and Multi-Scale Image Segmentations

Roger Trias-Sanz
Institut Géographique National; 2/4, avenue Pasteur, 94165 Saint-Mandé, France
SIP-CRIP5, Université René Descartes; 45, rue des Saints-Pères, 75006 Paris, France
Email: Roger.Trias.Sanz@ieee.org

Abstract— **We present a method for evaluating the quality of a multi-scale or hierarchical image segmentation against a reference single-scale segmentation, for comparing different segmentation algorithms or fine-tuning an algorithm's parameters to a specific application.**

I. Introduction

Image segmentation is usually performed as a preprocessing step for many image understanding applications, for example in some land-cover and land-use classification systems. A segmentation algorithm is used with the expectation that it will divide the image into semantically significant regions, or objects, to be recognized by further processing steps. It is however well known that semantically significant regions are found in an image at different scales of analysis. For a high-resolution aerial image, for example, at coarse scales we may find fields, while at finer scales we may find individual trees or plants. Parameters and thresholds in a typical single-scale segmentation algorithm must be tuned to the correct scale of analysis. However, it is often not possible to determine the correct scale of analysis in advance, because different kinds of images require different scales of analysis, and furthermore in many cases significant objects appear at different scales of analysis in the same image.

In an attempt to overcome this problem, in recent years there has been a trend toward multi-scale or hierarchical segmentation algorithms [1], [2]. These analyze the image at several different scales of analysis at the same time. Their output is not a single partition, but a hierarchy of regions, or some other data structure that captures different partitions for different scales of analysis.

As with classical, single-scale, segmentation algorithms, the need arises to evaluate the quality of a multi-scale segmentation against a reference, in order to compare different algorithms, and to select for an algorithm the parameters which are optimal for a given application. Most current segmentation evaluation methods [3], [4] handle only single-scale segmentations, that is, partitions of an image. They usually work by finding correspondences between points in the reference and points in the edges of the regions given by the segmentation. However, because multi-scale algorithms can deliver arbitrarily fine segmentations —at the finer end of the scale range— the concepts of "correspondence between reference points and segmentation edge points" and of "distance between segmentation edge and reference edge" cannot be easily transposed to the multi-scale case —in effect, at the right scale, the segmentation is so fine that all reference points are arbitrarily close to a segmentation edge.

We present a method for evaluating the quality of a multi-scale or hierarchical image segmentation against a reference. The reference segmentation is given as a set of edges of two kinds, *compulsory* and *optional*. The method computes two measures, a *false detection* (or comission) measure, and a *missed detection* (or omission) measure, by considering that a falsely detected segmentation edge is a bigger error if it appears at a coarse scale of analysis; conversely, the missed detection measure takes into account the fact that a reference edge can be completely missed —by there not being a corresponding segmentation edge— but also "nearly missed" if only fine-scale corresponding segmentation edges are found. This assumes that the most visually salient edges are found at the coarsest scales, which is the case. These two measures can be minimized jointly, or an aggregate such as their average can be calculated. We have found that they can be used to compare two different multi-scale segmentations of the same image.

This paper is structured as follows: In section II the quality measures are described in detail. Following that, section III presents examples of using this measure to study the behavior of a segmentation algorithm. Some concluding remarks are given in section IV.

II. Quality measures

The procedure to compute the segmentation quality measures operates by searching for pixels belonging to segmentation edges and for reference pixels. The first step is therefore to convert the inputs into a suitable discrete (pixel based) form.

The segmentation reference is given as two sets of segments, one for the compulsory edges and one for the optional edges. A rasterization algorithm —such as Bresenham's line drawing method— is applied to convert these sets of edges into two sets of pixels. Note that information about what edge each pixel belongs to is lost. Let $D \subset \mathbb{Z}^2$ be the usually rectangular image domain. Pixels coordinates are elements of D. Let $R_c \subset D$ be the set of pixels given by the compulsory reference edges. Let $R_o \subset D$ be the set of pixels given by the optional reference edges, and $R = R_c \cup R_o$.

The hierarchical or multi-scale segmentation must also be converted into a flat, discrete, representation. Guigues [1] and others suggest that multi-scale segmentations should be *causal*,

0-7803-9051-2/05/$20.00 (C) 2005 IEEE

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
25 JUL 2005

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
A Metric for Evaluating and Comparing Hierarchical and Multi-Scale Image Segmentations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institut Géographique National; 2/4, avenue Pasteur, 94165 Saint-Mandé, France

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001850, 2005 IEEE International Geoscience and Remote Sensing Symposium Proceedings (25th) (IGARSS 2005) Held in Seoul, Korea on 25-29 July 2005. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**4**	

i.e., segmentation edges found at coarser scales must have corresponding edges at finer scales. When corresponding finer edges have exactly the same position as the coarser edges, as is the case in Guigues' algorithm, the flattening technique described in [5] can be used: Each edge $e \in E$ can be found in segmentations at a given coarsest scale $\lambda_E^+(e)$ and at all finer scales. The edge e is rasterized and converted to a set of pixels, $p(e) \subset D$. Let $S = \cup_{e \in E} p(e)$. For each pixel s in S, its coarsest scale is computed as

$$\lambda_S^+(s) = \max\{\lambda_E^+(e) : s \in p(e), e \in E\}. \tag{1}$$

However, the behaviour of $\lambda_E^+(e)$ is highly dependent on the specific segmentation technique and parameter set used. To avoid this, we use not $\lambda_S^+(s)$ but this transformation: All values of λ_S^+ are sorted in decreasing order, keeping repeated values. Then each is replaced by the exponential of its rank —the rank being 0 for the largest value of λ_S^+ and $|S|-1$ for the smallest; repeated values are given the rank of their first occurence— as

$$w^+(s) = \mathrm{e}^{-\alpha \cdot \mathrm{rank}(\lambda_S^+(s))}, \tag{2}$$

where α is a user-defined parameter.

In other causal algorithms where the finer corresponding edge can have a different position from that of the coarser edge, each edge e can be seen to keep its position between scales $\lambda_E^-(e)$ and $\lambda_E^+(e)$. We can then correspondingly define λ_S^-, w^-, and, in the following discussion, use $w = w^+ - w^-$ instead of w^+.

Computing the missed detection and false detection measures involves searching in small circular neighborhoods of pixels, of fixed radius ϵ. The neighborhood radius ϵ is used as a distance threshold when determining if a segmentation pixel and a reference pixel could correspond and, therefore, it must be set to take into account positional accuracy in the reference and acceptable positional errors in the segmentation edges, and must keep the same value for all tests in a comparison of segmentations.

Given the R and R_c reference sets, the S segmentation set, the weight w^+ map, and the neighborhood radius ϵ, the two quality measures are defined as follows:

Let B_x be the ball in D centered on x with radius ϵ,

$$B_x = \{y \in D : \|y - x\| \leq \epsilon\}. \tag{3}$$

The missed detection penalty for a single pixel $x \in R_c$ is defined as

$$p_m(x) = \begin{cases} 1 - \max\limits_{y \in B_x \cap S} w^+(y) & \text{if } B_x \cap S \neq \emptyset, \\ 1 & \text{if } B_x \cap S = \emptyset, \end{cases} \tag{4}$$

and the false detection penalty for a single pixel $x \in S$ as

$$p_f(x) = \begin{cases} 0 & \text{if } B_x \cap R \neq \emptyset, \\ w^+(x) & \text{if } B_x \cap R = \emptyset. \end{cases} \tag{5}$$

The global missed detection quality measure is

$$M = \frac{\sum_{x \in R_c} p_m(x)}{\operatorname{card} R_c}, \tag{6}$$

and the global false detection quality measure is

$$F = \frac{\sum_{x \in S} p_f(x)}{\sum_{x \in S} w^+(x)}. \tag{7}$$

The missed detection —or omission— penalty is computed for all pixels corresponding to compulsory ground truth segmentation edges. The penalty is maximum, as in single-scale segmentations, when no corresponding segmentation edge pixel is found. However, a penalty is also given when a corresponding pixel is found, and is lower as the pixel's scale is higher. Therefore, if the only segmentation pixel corresponding to a ground truth edge is found at very fine scales of analysis, it will be counted as an "almost-missed detection". On the contrary, if the corresponding pixel is very salient and appears at coarse scales of analysis, the penalty will be minimal. Conversely, the false detection or commission penalty is computed for all pixels corresponding to detected segmentation edges, that is, edges present in the segmentation algorithm's output. No penalty is given if a corresponding ground truth edge is found. If none is found, the pixel's scale is used for the penalty. Therefore, very salient detected edges that do no correspond to the ground truth have a high penalty, while incorrectly detected edges at fine scales of analysis are more tolerated.

This provides a two-dimensional measure of the quality of a segmentation compared to a ground truth, $(M, F) \in [0,1] \times [0,1]$. The closer M and F are to zero, the higher the quality. These measures can be used to compare different segmentation algorithms or parameter sets, but should not be taken to be absolute quality measures —that is to say, a value of $F = 0.5$ does not mean that half the detected edges are false detections. In order to find the optimal parameter set these two measures can be minimized jointly, using the partial order

$$(M_1, F_1) \leq (M_2, F_2) \iff M_1 \leq M_2 \cap F_1 \leq F_2, \tag{8}$$

or an appropriate aggregate such as their average can be calculated and that scalar value be minimized.

III. Examples

To demonstrate the presented quality measures, in this section we will show how they vary as parameters in a multiscale segmentation algorithm are modified. Guigues' hierarchical segmentation method [1] will be used; however, it is not the goal of this paper to thoroughly evaluate the merits of the algorithm itself or of specific parameter sets, but to show that the quality measures presented in this paper can be used to do so in a systematic way.

In all graphs shown in this section, the horizontal axis corresponds to the "missed detection" measure M, and the vertical axis to the "false detection" measure F. The closer to the origin the measures, the higher the segmentation quality.

Fig. 1 and Fig. 2 show how the quality measures change as we randomly modify a segmentation. In Fig. 1, random values following a Gaussian law with increasing variance were added to the segmentation weights $w^+(s)$ before evaluation. In Fig. 2, the image segmentation was geometrically distorted by adding a random offset —following a Gaussian law with increasing variance— to the positions of endpoints of segmentation edges E.

Fig. 3 shows the effect of using a feature of Guigues' segmentation algorithm that gives a higher penalty to segmentation edges not following image pixels of strong gradient module —the basic Guigues' segmentation algorithm is region-based and uses the distribution of pixel values within each segmentation region, and the shape of the region's boundary, but not the gradient at the region's boundary. It can be seen that making the contribution of that penalty smaller or larger has a very limited effect on the resulting segmentation, suggesting that this gradient-following penalty is actually redundant with the basic Guigues method. The close-up in Fig. 4, where data points obtained by increasing the contribution of this gradient-based penalty are joined by lines, confirms that this penalty produces only small random variations.

In contrast, Fig. 5 shows the effect of using different image channels for the segmentation algorithm. Several tests are shown, using, for example, the raw RGB channels, derived color spaces (HSV, log-opponent chroma, CIE, ...) and textural features (gradient direction entropy, space orientation histograms, ...). It can be seen that the choice of input channels has a much more significant effect on the resulting segmentation.

Fig. 6 shows two of the segmentations used in the graph in Fig. 5. The source image is at the top; the segmentation at the middle uses the value and hue components (combined in polar form), and the CIE A and CIE U channels as input channels, and corresponds to the data point shown with a large circle in Fig. 5. The segmentation at the bottom uses a feature based on Zhou's scale-orientation histograms [6] (SOHs) and the entropy of the local histograms of gradient directions [7] as input channels, and corresponds to the data point shown with a large triangle in Fig. 5. The first segmentation, which is clearly better by visual analysis, has correspondingly better results for both the M and the F measures.

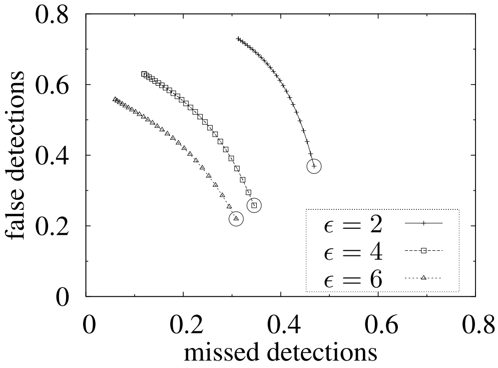

Fig. 1. Adding Gaussian perturbations to segmentation weights $w^+(s)$, for several values of ϵ and increasing perturbation variance. The larger circles show the results with no perturbation.

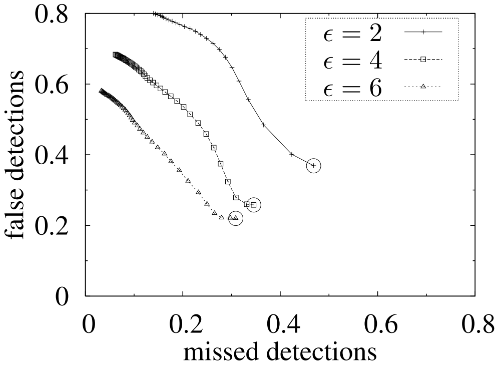

Fig. 2. Adding Gaussian offsets to the positions of segmentation edge endpoints, for several values of ϵ and increasing perturbation variance. The larger circles show the results with no perturbation.

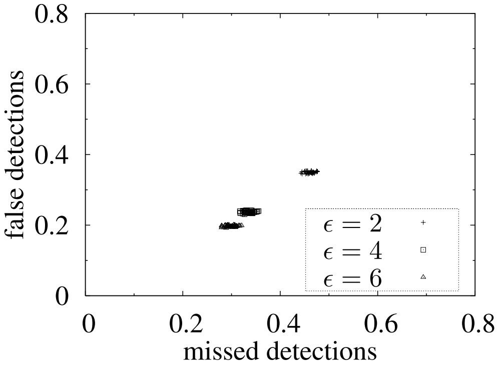

Fig. 3. Modifying the weight of the gradient-following penalty, for several values ϵ.

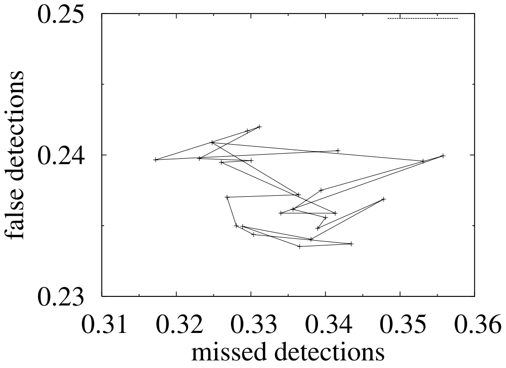

Fig. 4. Modifying the weight of the gradient-following penalty, close-up for $\epsilon = 4$.

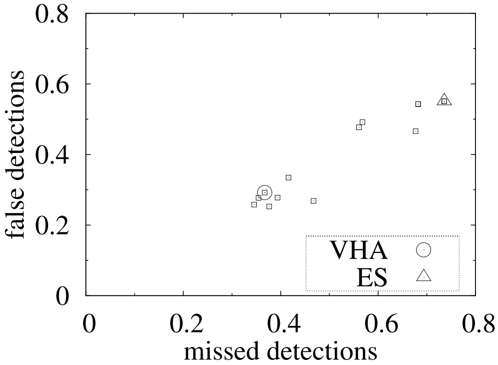

Fig. 5. Using different color space components and textural features as input to the segmentation. Shown for $\epsilon = 4$. The larger circle ("VHA") and triangle ("ES") mark the segmentations that are shown in Fig. 6.

IV. Conclusion

Multi-scale approaches have been used for image segmentation for some time to avoid the problem of having to select a globally valid scale of analysis before actual image interpretation. In order to evaluate multi-scale image segmentation algorithms and their parameter sets, a method is needed to compare a segmentation to a ground truth. However, current evaluation methods are not designed for use in multi-scale segmentations. In this paper we propose a two-dimensional quality measure that evaluates the comission and omission errors of a multi-scale segmentation against a single-scale ground truth. We show several examples of how these measures behave with varying segmentation parameter sets.

References

[1] L. Guigues, H. Le Men, and J.-P. Cocquerez, "Scale-sets image analysis," in *Proc. IEEE Intl. Conf. on Image Processing (ICIP 2003)*. Barcelona, Spain: IEEE, Sept. 2003.

[2] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 561–576, Apr. 2000.

[3] M. Segui Prieto and A. R. Allen, "A similarity metric for edge images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1265–1272, Oct. 2003.

[4] J. Liu and Y.-H. Yang, "Multiresolution color image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 7, pp. 689–700, 1994.

[5] R. Trias-Sanz, "An edge-based method for registering a graph onto an image with application to cadastre registration," in *Proc. of the 2004 Conf. on Advanced Concepts for Intelligent Vision Systems (ACIVS 2004)*, Brussels, Belgium, 2004, pp. 333–340.

[6] J. Zhou, L. Xin, and D. Zhang, "Scale-orientation histogram for texture image retrieval," *Pattern Recognition*, vol. 36, pp. 1061–1062, 2003.

[7] C. Baillard, "Analyse d'images aériennes stéréo pour la restitution 3-d en milieu urbain," Ph.D. dissertation, École Nationale Supérieure des Télécommunications, Paris, France, Oct. 1997, (In French).

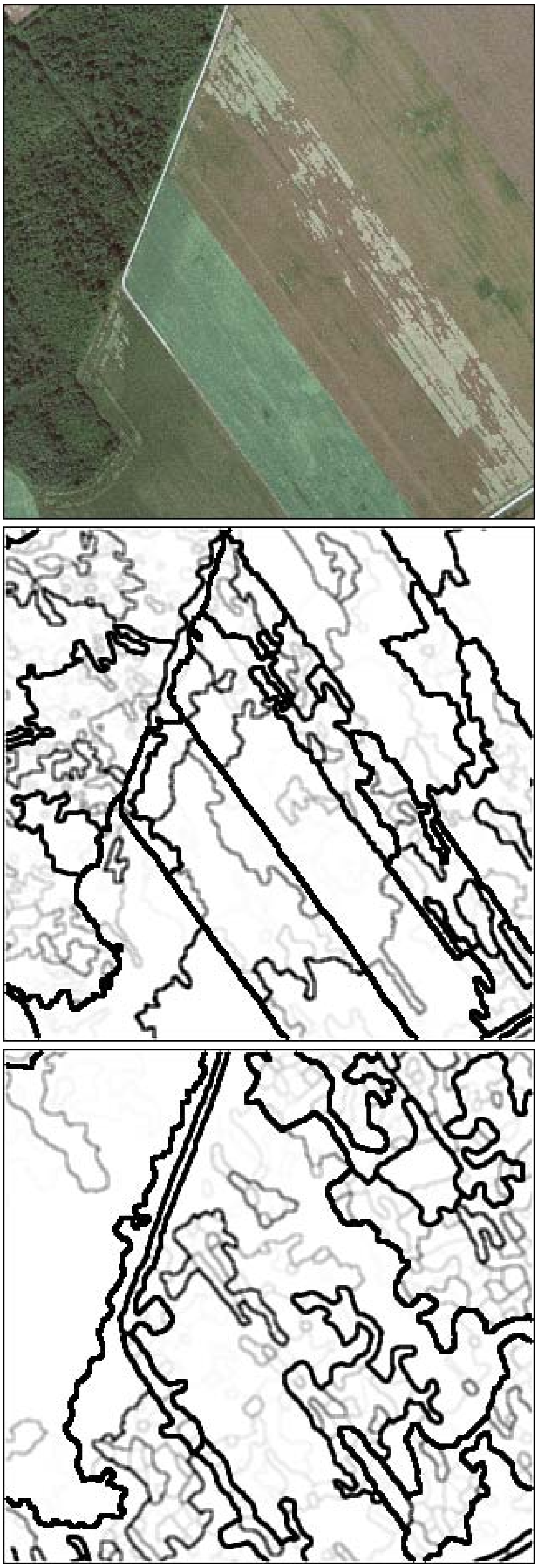

Fig. 6. Using different color space components and textural features as input to the segmentation. Top: source image. Middle: segmentation with value, hue, CIE A and CIE U channels. Bottom: segmentation with SOHs and direction histogram entropy. Segmentations are shown by their w^+ image, with darker pixels corresponding to higher values of $w^+(s)$.